Comparison of Classification Methods for Spectral Data of Laser-induced Fluorescence

Marian Kraus[1], Lea Fellner[1], Florian Gebert[1], Karin Grünewald[1], Carsten Pargmann[1], Arne Walter[1], and Frank Duschek

German Aerospace Center, Institute of Technical Physics, Langer Grund, Lampoldshausen, 74239 Hardthausen, Germany,
marian.kraus@dlr.de,
WWW home page: http://dlr.de/tp/en

Abstract. Online detection of CBRNE is a research field of growing importance due to its relevance for public security and defense. The selectivity of machine learning has reached maturity in order to distinguish very similar laser-induced fluorescence (LIF) spectra of different samples - establishing the basis for an automatic classification. The work in this contribution applies the classification process of decision trees, support vector machines and artificial neural networks to LIF spectra. Two experimental setups with two excitation wavelengths each (280 and 355 nm; 266 and 355 nm) and different spectral resolutions of about 1 nm and 12 nm, respectively, have been performed. In the first setup the discrimination of seven bacteria species with an accuracy of over 90 % is demonstrated. The data of the second setup with lower spectral resolution are equally sufficient for a subsequent classification. The results are compared and represented in a low-dimensional subspace for the purpose of visualization.

Keywords: Standoff detection; Laser-induced fluorescence spectra; Classification models; Machine learning; Decision trees; Support vector machines; Artificial neural networks

1 Introduction

A fast detection of hazardous substances can save human lives. Many techniques have been developed, more are in progress and the demand is still growing. The setups vary widely and so do the subsequent classification methods [1–3]. This paper compares the performance of three different algorithms applied to two datasets, collected with two experimental setups, in which LIF spectra of seven bacterial samples are discriminated. On the one hand their is a noisy signal, on the other hand their is a lower resolution, but the classification procedure is similar.

For investigations in this paper the following bacterial species are used: *Bacillus subtilis, Bacillus thuringiensis, Brevibacillus brevis, Escherichia coli K12, Micrococcus luteus, Oligella urethralis and Paenibacillus polymyxa.* For each

bacterial strain the measurements performed with both setups were made with samples originating from the same cultures and the suspensions were measured within few hours after preperation.

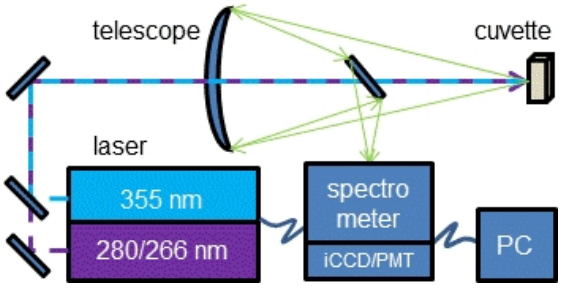

Fig. 1: The spectra were recorded with two different setups using two excitation wavelengths, each, both illustrated in this simplified schematic view.

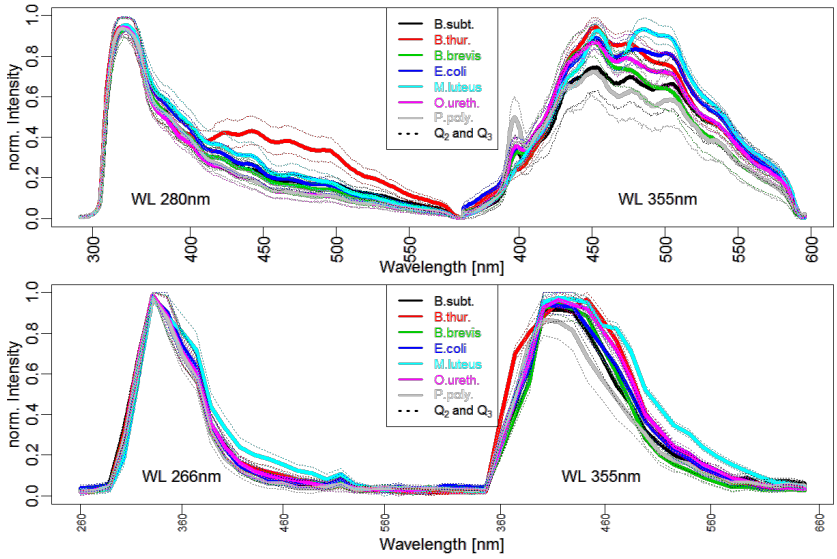

Fig. 2: The resulting signals of both setups: reduced to important ranges, rescaled and corresponding data merged. The dotted lines enclose 50 % of the spectra (2^{nd} and 3^{rd} quartile).

The first setup utilizes two alternating laser pulses with excitation wavelengths of 280 and 355 nm. The cuvettes are placed outdoors at a distance of 22 m and the fluorescence signal is recorded by an intensified CCD camera with 1024 channels [1]. The second setup is based on a simultaneous emission of two laser pulses with wavelengths of 266 and 355 nm, the latter temporally shifted by an optical delay line. In this case the samples are placed indoors at a distance

of 3.5 m and the signal is gathered by a photomultiplier tube array with 32 channels [4]. The experimental setups (figure 1) and data acquisition are fully described in ref. [1] and the proceedings of *SICC 2017* [4]. Figure 2 shows the spectra after the preprocess described in section 3.1.

For statistical computing the free software environment R [5] is used, especially the *caret* package [6]. This combines classification packages, i.a. ref. [7–11], which were used individually for some illustrations of this contribution. A detailed description of the algorithms and their mathematical background can be found in ref. [12] and [13].

2 Classification Algorithms

2.1 Decision Tree

A decision tree is used to divide sets in several subsets until there is no more splitting needed to accomplish a distinct classification. The investigation of single features (channels) leads to a couple of benchmarks to make a binary decision at each node saying if feature X is bigger than a value Y go to the next node A, otherwise go to node B. Following these instructions each single dataset is guided along the branches until it is associated at the leafs to a specific class.

The model construction is done in multiple steps where the tree is rearranged based on the information entropy of every channel, like single features that discriminate a whole class from the rest. In addition, the placement of a border separating two classes is varied within the gap between. The one which has the least mean squared distance to both classes is used. The result is a more effective model regarding to its size and evaluation time. Figure 3 shows a small model which classifies the data in a few steps observing five features from setup 2.

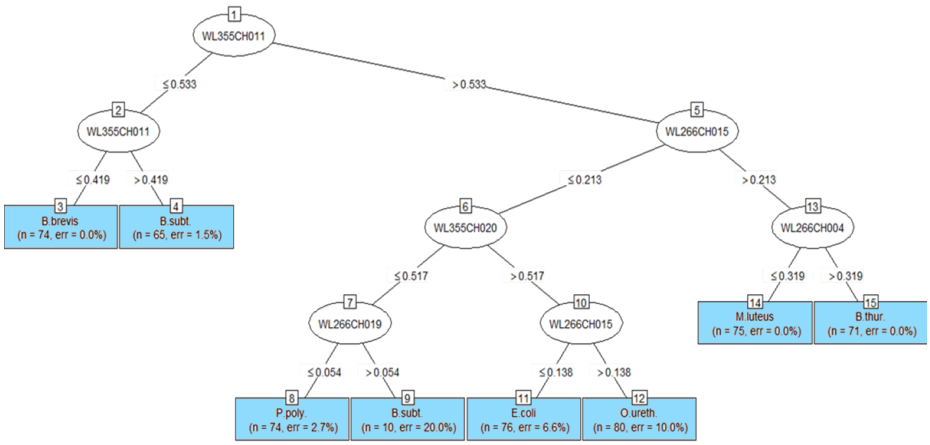

Fig. 3: Decision Tree example demonstrating the classification principle

2.2 Support Vector Machine (SVM)

Decision trees imply the ability of linear separation but in many times this is not possible. One can make an additional transformation of the data by mapping them into a hyperspace (commonly with the radial basis function [13]). This procedure is called *Kernel Trick* and enables a linear division in higher dimensions. The closest points to the boundaries are the *support vectors* and only they take affect on the border's shape (see figure 4).

Figure 5 shows a binary classification with only two features and *Bacillus thuringiensis* is the one we want to detect. The model is defined just by the *support vectors* marked with filled black symbols. Many of the spectra are classified correctly using only two detection channels.

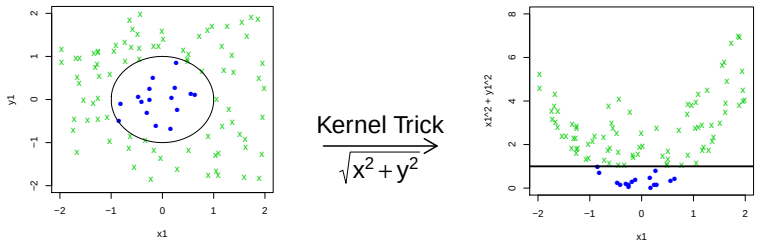

Fig. 4: This example of a SVM classification divides two classes using the distance from the origin as kernel function. After the transformation a simple linear separation is possible.

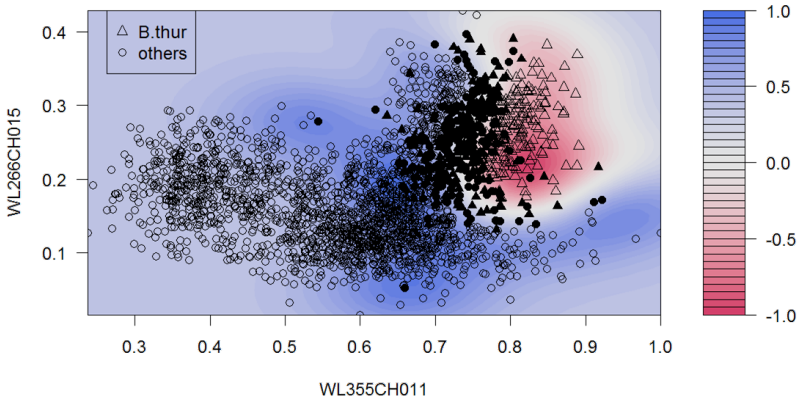

Fig. 5: This SVM model shows the classification of one sample versus the others focusing on only two features. A combination with all pairs of channels generates the model.

2.3 Artificial Neural Network (ANN)

Classifying is a function, mapping the feature space (*input* I) onto a 'class space' (*output* O). ANNs use two functions: a linear combination maps features to an additional space (*hidden units* H), then mapping this to the classes. Connections are weighted by the coefficients of the linear function, changing with every iteration step, influenced by the coefficients of the previous second mapping.

Figure 6 shows an 10-7-7 ANN: I are ten features with large variance, H is one hidden layer with seven variables and O are the seven species. The bias terms B with a constant influence represent simulated noise. The amount of inputs and hidden units can be varied as well as the maximum of iteration steps.

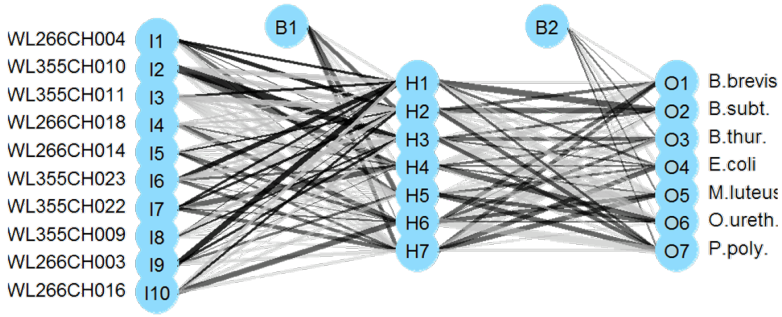

Fig. 6: This ANN has ten features and seven hidden units showing the weights after 1000 iterations. Dark connections represent positive, bright ones negative coefficients and the widths represent the absolute values of the coefficients for the linear combination.

3 Data Processing

3.1 Preprocess

After background correction and outlier removal each spectrum of setup 1 was smoothed by running median due to a lower signal to noise ratio. Features were eliminated which cannot yield useful information like channels beyond the lower wavelength, where no fluorescence can occur, or the filter range at about 355 nm. The Raman peak of water, still shown on the right side of figure 2 at about 400 nm, was removed for further operations because of its misleading intensity. After rescaling every single spectrum the corresponding data of both wavelengths were concatenated to construct one dataset for the classification process.

Each bacterial sample was measured five times gathering 100 spectra per measurement and wavelength. Thus we obtained a total set of 3500 records per setup. The upper half of figure 2 displays the averaged spectra acquired with

setup 1 and the boundaries of the second and third quartile containing 50 % of each class. The lower part visualizes the data of the second setup accordingly. Detailed descriptions of the spectra are given in ref. [1] and [4].

3.2 Training and Test

The data were randomly split in two parts. For reasons of verification repeated by resampling methods like cross validation and bootstrapping. 75 % were used to train the algorithms with different parameters to find those with the highest *accuracy*, ratio of correctly classified spectra and the total. The remaining 25 % were assigned with the generated models to test their goodness of fit.

To prevent overfitting the training parameters are set not too exactly. For example the minimal number of spectra in the leafs of a decision tree could be too small and so the algorithm tries to match even those terms which are noisy and only therefore belong to the area of another class. This will lead to misclassifications applying the model on unknown data.

4 Results and Discussion

The three presented algorithms discriminated seven bacteria samples better than 92 % only by observing their LIF spectra. Table 1 shows the performance for the test datasets from both setups and the three different methods from above. Decision trees gained an *accuracy* of 92.1 % and 97.7 %, dependent on the setup. The SVM obtained the best results of 96.5 % and 99.5 %, respectively. Also the ANN provided a very good classification (*accuracy*: 92.6 % and 99.1 %).

The rather noisy signal of setup 1 and the low resolution of setup 2, both yield enough information for different methods of machine learning. The same data can be used to generate other models with even better performance depending on the preprocess as well as on tuning parameters for each method.

Due to the lower dimensionality of the data acquired by setup 2 the training process is much faster without using any type of feature selection. Instead of some hours it only takes several minutes and allows a wider search for optimal tuning parameters in the same runtime.

Grouping the bacteria in two classes like 'harmless' and 'harmfull' and using the same models the *accuracy* could not be worse. There are no new misclassifications but some of the so-called *false positives* and *false negatives* now belong to the same group and are correct. Therefore, the performance values would increase compared to the previous prediction.

5 Conclusion and Outlook

The results show that it is possible to distinguish between seven different bacteria by analyzing single spectra of LIF. An established technique with unexhausted possibilities in standoff detection of organisms due to the presence of special amino acids.

Latest studies of bacteria in different growth phases show the changing characteristics of their LIF spectra [14]. This behavior has to be taken into account during the process of model generation.

The preprocess could include more steps like principle component analysis or feature selection. This reduction would lead to a faster model generation and less overfitting and will be part of future examinations. Taking the average of five consecutive spectra would obtain even better values but we aimed for a single shot classification.

It is promising that a combination of at least three models would increase the validity and help to eliminate samples which were mismatched in a minority of cases. Investigating other algorithms, e.g. random forests, a pool of models could be generated, each assigning the spectra for its own. Having used every model an ambiguity may still exist which is solved by an *hybrid* model choosing the classification appearing more often.

References

1. Frank Duschek, et al. *Standoff Detection and Classification of Bacteria by Multispectral Laser-Induced Fluorescence. Advanced Optical Technologies*, 6(2):75–83, April 2017.
2. Pasqualino Gaudio, et al. *Application of optical techniques to detect chemical and biological agents. Defence S&T Technical Bulletin*, 10(1):1–13, 2017.
3. Sylvie Buteau, et al. *Laser based standoff detection of biological agents.* Technical Report TR-SET-098, North Atlantic Treaty Organisation, Research and Technology Organisation, 2010.
4. Florian Gebert, et al. *Standoff detection and classification of chemical and biological hazardous substances combining temporal and spectral laser induced fluorescence techniques.* In *First Scientific International Conference on CBRNe.* Scientific International Conference on CBRNe, May 2017.
5. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.
6. Max Kuhn. *caret: Classification and Regression Training*, 2017. R package version 6.0-76.
7. Max Kuhn, et al. *C50: C5.0 Decision Trees and Rule-Based Models*, 2015. R package version 0.1.0-24.
8. Alexandros Karatzoglou, et al. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.
9. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
10. Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2016. R package version 1.12.0.
11. David Meyer, et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015. R package version 1.6-7.
12. Brett Lantz. *Machine Learning with R - Second Edition.* Packt Publishing, 2015.
13. Max Kuhn and Kjell Johnson. *Applied Predictive Modeling.* Springer, 2013.
14. Lea Fellner, et al. *Variations in fluorescence spectra of a bacterial population during different growth phases.* In *1st Scientific International Conference on CBRNe.* Springer, 2017.

Table 1: Correlation matrices for both setups and three classification models: the six diagonals are constituted by the correctly classified spectra

		Setup 1							Setup 2						
Classified as →		*B.brevis*	*B.subt.*	*B.thur.*	*E.coli*	*M.luteus*	*O.ureth.*	*P.poly.*	*B.brevis*	*B.subt.*	*B.thur.*	*E.coli*	*M.luteus*	*O.ureth.*	*P.poly.*
Decision Tree Algorithm															
					92.1 %			*Accuracy*				97.7 %			
actual sample	*B.brevis*	114	3						125						
	B.subt.		102				1	14	2	122					1
	B.thur.			123	1		1				124			1	
	E.coli	2	2		114		6	1	1	1		119			4
	M.luteus					123	1	1					125		
	O.ureth.	2			3	1	119				1	3	1	120	
	P.poly.	5	6		2		1	111		2		2		1	120
Support Vector Machine															
					96.5 %			*Accuracy*				99.5 %			
actual sample	*B.brevis*	120	1				1	3	125						
	B.subt.		119					6	1	124					
	B.thur.			124			1				125				
	E.coli				121		4					124		1	
	M.luteus					123		2					125		
	O.ureth.	2					123							125	
	P.poly.	4	2		5			114				2			123
Artificial Neural Network															
					92.6 %			*Accuracy*				99.1 %			
actual sample	*B.brevis*	114	4				5	2	125						
	B.subt.	2	112		2	2	2	5	1	124					
	B.thur.			125							125				
	E.coli	3			113		5	4				123		1	1
	M.luteus	1	2			121		1					125		
	O.ureth.	1			3		120	1				2		123	
	P.poly.	6	8		4		2	105		2		1			122